Biased Distribution of Inverted and Direct *Alu*s in the Human Genome: Implications for Insertion, Exclusion, and Genome Stability

Judith E. Stenger,[1,4] Kirill S. Lobachev,[2] Dmitry Gordenin,[2] Thomas A. Darden,[1] Jerzy Jurka,[3] and Michael A. Resnick[2,5]

[1]*Laboratory of Structural Biology,* [2]*Laboratory of Molecular Genetics, National Institute for Environmental Health Sciences, NIH, Research Triangle Park, North Carolina 27709, USA;* [3]*Genetic Information Research Institute, Sunnyvale, California 94089, USA*

Alu sequences, the most abundant class of large dispersed DNA repeats in human chromosomes, contribute to human genome dynamics. Recently we reported that long inverted repeats, including human *Alu*s, can be strong initiators of genetic change in yeast. We proposed that the potential for interactions between adjacent, closely related *Alu*s would influence their stability and this would be reflected in their distribution. We have undertaken an extensive computational analysis of all *Alu*s (the database is at http://dir.niehs.nih.gov/ALU) to better understand their distribution and circumstances under which *Alu* sequences might affect genome stability. *Alu*s separated by <650 bp were categorized according to orientation, length of regions sharing high sequence identity, distance between highly identical regions, and extent of sequence identity. Nearly 50% of all *Alu* pairs have long alignable regions (>275 bp), corresponding to nearly full-length *Alu*s, regardless of orientation. There are dramatic differences in the distributions and character of *Alu* pairs with closely spaced, nearly identical regions. For *Alu* pairs that are directly repetitive, ~30% have highly identical regions separated by <20 bp, but only when the alignments correspond to near full-size or half-size *Alu*s. The opposite is found for the distribution of inverted repeats: *Alu* pairs with aligned regions separated by <20 bp are rare. Furthermore, closely spaced direct and inverted *Alu*s differ in their truncation patterns, suggesting differences in the mechanisms of insertion. At larger distances, the direct and inverted *Alu* pairs have similar distributions. We propose that sequence identity, orientation, and distance are important factors determining insertion of adjacent *Alu*s, the frequency and spectrum of *Alu*-associated changes in the genome, and the contribution of *Alu* pairs to genome instability. Based on results in model systems and the present analysis, closely spaced inverted *Alu* pairs with long regions of alignment are likely at-risk motifs (ARMs) for genome instability.

The genomes of many complex organisms contain short, interspersed, intermediate repetitive elements that are nonviral, nonautonomous transposons. The *Alu* sequence elements, which are derived from 7sRNAs, are the most numerous in primates (for review, see Novick et al. 1996), with the human genome containing over 1,000,000 *Alu* sequences, or ~10% of the total DNA (Jurka 1998).

The dissemination of diverged *Alu* repeats over the last 65 million years have contributed to the structure, function, evolution, and diversity of the human genome. *Alu*s have been regarded as "junk" DNA because of their high frequency and inert nature. Retroposed *Alu* insertions, however, can coevolve in the context of their target DNA and hence take on diverse functions (Szmulewicz et al. 1998). Besides becoming a structural genetic component by integrating into an exon (Mullersman and Pfeffer 1995), *Alu* elements can serve in a regulatory capacity through diverse mechanisms. They also can behave as a modulator of DNA replication (Tsuchiya et al. 1998), a positive (Norris et al. 1995; Gu et al. 1997) or negative enhancer, a regulator of an enhancer, a mediator of alternative splicing, or have a role in genetic imprinting. Recently, *Alu*s also have been proposed to downregulate translation in response to cellular stress and viral infection by antagonizing double-stranded RNA-activated kinase PKR activation (Chu et al. 1998) and they are the subject of p53 transcriptional control (Chesnokov et al. 1996). These many *Alu* functions help to explain the retention of *Alu* sequences in the genome.

*Alu*s also may have negative consequences and impact on human health. Besides their ability to retropose to inappropriate regions or to facilitate unequal homologous recombination events (Deininger and Batzer 1999), *Alu* sequences consist of highly methylated CpG islands that potentiate a risk for point mutations through deamination of 5′ methyl-deoxycytidine. *Alu* insertions have been found in the vicinity of several diverse human disease-associated genes and

[4]Present address: Duke Center for Human Genetics, Duke University Medical Center, Box 3445, Durham, NC 27710, USA.
[5]Corresponding author.
E-MAIL resnick@niehs.nih.gov; FAX (919) 541-7593.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.158801.

11:12–27 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

Alu-mediated mutagenesis has been estimated to contribute to 0.4% of new human genetic diseases (Deininger and Batzer 1999). Examples include breast cancer (Miki et al. 1996), acute myeloid leukemia (Strout et al. 1998), neurofibromatosis (Wallace et al. 1991), hemophilia B (Vidaud et al. 1993), hypertension (Muratani et al. 1993; Kitamura et al. 1996; Anderson et al. 1998), Ehlers-Danlos syndrome type VI (Pousi et al. 1994, 1998; Toriello et al. 1996; Heikkinen et al. 1997), *alpha-zero-thalassemia* (Harteveld et al. 1997), and Tay-Sachs disease (Myerowitz and Hogikyan 1987).

Based on their abundance and an average of approximately 85% identity (Shen et al. 1991), they are considered as potential sites of recombination and, therefore, threats to genome stability. Although several diseases have been associated with rearrangements involving *Alu* interactions including cancer (Miki et al. 1996; Slebos et al. 1998; Strout et al. 1998), Tay-Sachs disease (Myerowitz and Hogikyan 1987), and hypercholesterolemia (Chae et al. 1997), *Alu*s generally appear to be relatively stable. However, they have the potential for instability. For example, high levels of recombination between artificial direct identical repeats have been observed in p53-defective human cells (Gebow et al. 2000).

What determines whether *Alu* elements will be benign or pose a threat to human health? Aside from their appearance within important regions of critical genes, we have pursued features of *Alu* elements and their arrangement in the human genome that would identify potential destabilizing effects as well as suggest their mode(s) of integration and/or stability. One approach is to examine the pairwise distribution of *Alu*s with the idea that adjacent *Alu*s might interact with higher frequency than *Alu*s that are farther apart. *A priori*, the appearance and characteristics of an *Alu* could be independent of other *Alu*s or, alternatively, might be influenced by adjacent *Alu*s. For example, a nonrandom, high frequency of closely spaced *Alu*s might indicate preferential insertion. Therefore, a study of *Alu* pairs may reveal mechanisms of insertion and/or subsequent preferred changes. Findings from previous computational analysis of *Alu* distributions suggest a strong bias towards pairs in which the *Alu*s are in a direct orientation and closely spaced (Jurka 1995). They appear to integrate preferentially into AT-rich regions, for example, at the poly(A) tails of preexisting *Alu*s. There also is a nonrandom differential distribution of *Alu* repeats, with clustering in some regions and a paucity of sequences in others.

Driven in part by questions of homologous interactions, we have developed approaches to analyzing *Alu* pairs based on regions of alignments and degree of sequence identity. Observations with yeast and mouse cells suggest that long closely related inverted repeats are unstable and can cause deletions in eukaryotes (Gordenin et al. 1993; Ruskin and Fink 1993; Lewis 1999; Lewis et al. 1999). Moreover, they are initiators of recombination in yeast (Gordenin et al. 1993; Nag and Kurst 1997; Lobachev et al. 1998). This has led us to propose that long inverted repeats can be at-risk motifs (ARMs) in the genome (Gordenin and Resnick 1998).

Recently, we demonstrated with a yeast-based model system that inverted *Alu* pairs are hotspots for recombination, even if diverged (Lobachev et al. 2000), suggesting that they could be sites of genomic instability in the human genome. A preliminary analysis of adjacent full-length *Alu*s revealed that there were relatively few closely spaced inverted *Alu*s as expected based on the results in yeast (Lobachev et al. 2000). This has led to the present extensive analysis of the distribution of *Alu* pairs in the human genome in order to identify motifs that are excluded and to describe motifs that might be at risk. Having established several factors that could influence the ability of inverted *Alu* pairs to initiate genetic change in yeast, we examined computationally the distribution of *Alu* pairs in the human genome according to the orientation of *Alu*s within a pair, length of aligned regions, identity of aligned regions, distance between aligned regions, and age. While full-length *Alu*s are ~300 bp, they frequently are truncated. We found that the *Alu*s in nearly half of all pairs share long regions (>275 bp) of alignment regardless of orientation and that there is a bias for young *Alu*s in pairs whose *Alu*s are separated by <40 bp. There are several differences in the distribution of direct and inverted *Alu* pairs. The aligned regions of direct *Alu* pairs are frequently separated by <20 bp when the alignments correspond to full-size or half-size *Alu*s. *Alu* pairs with long regions of alignment (>275 bp) also tended to be more closely related than shorter *Alu* pairs. Unlike direct *Alu* repeats, inverted repeats with closely spaced (<20 bp) aligned regions are rare, suggesting they are excluded from formation and/or are unstable once formed. The inverted and direct *Alu* pairs also exhibit very different patterns of truncation that likely reflect mechanisms of integration. Based on results in yeast and the observed distributions of adjacent *Alu* repeats, we suggest that there are inverted *Alu* motifs that are potentially unstable and may be sources of chromosome instability in somatic tissue.

RESULTS

Approaches to Analyzing *Alu* Pair Distribution

Our analysis of *Alu* pairs was guided by observations in model systems that investigated stability of large repeats and their potential for interaction. In particular, results with human *Alu*s and other long repeats in yeast (Nag and Kurst 1997; Lobachev et al. 1998, 2000)

and other organisms (for review, see Erlich 1989; Leach 1994) suggested that orientation, distance between regions of identity, degree of identity, and length of highly related regions could influence the likelihood that adjacent *Alu* sequences might impact on genome stability. Biases in various classes of *Alu* repeats were expected to indicate differences in targeting and/or stability of *Alu* pairs.

First, we identified *Alu* sequences by comparing them to a well-defined consensus sequence (see Methods). *Alu* pairs were identified in which the separation was <650 bp and the orientations, direct and inverted, were identified (see Methods). Briefly, *Alu*s were assigned an orientation D or C according to their direction in the human genome sequence database: direct with the poly(A) tail at the 3′ end or complementary with the complement of the tail at the 5′ end. There are four possible orientations for the *Alu* pairs corresponding two combinations of direct repeats (DD and CC) and two combinations of inverted repeats (CD with the double-strand AT-rich tails pointed outward and DC with AT tails pointed inward). In addition to orientation, the pairs were analyzed according to degree of homology (defined by the identity score; see Methods) and length of identical regions across the aligned *Alu* region ("a" in Fig. 1). Both the distance between *Alu* sequences (c) and more importantly the distance between aligned regions in *Alu* pairs were determined (see Fig. 1 and Methods). Based on the results of Lobachev et al. (2000), the latter would be an important parameter in determining the potential for instability. The distance between aligned regions was equal to the unique *spacer* sequence "c" between two *Alu*s plus any *Alu* sequence "b" between them that was not common to both *Alu*s (i.e., there was a truncation or possibly a highly diverged region). Thus, the actual distance between aligned regions within *Alu* pairs (b + c) could be much longer than the apparent distance between *Alu*s; however, this was atypical.

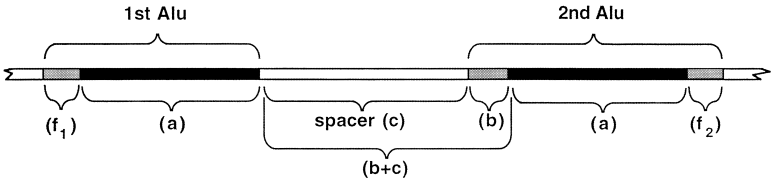

Figure 1 Diagram of an adjacent pair of *Alu*s with aligned regions. Two *Alu* sequences are presented that can be in inverted or direct-repeat orientation. The *Alu* sequences are identified by virtue of sharing homology with an *Alu* consensus sequence (see Methods). The alignment length (a) corresponds to a common (although nonidentical) region shared by both. External flanking sequences (f1) and (f2), if present, correspond to *Alu* regions of no (i.e., a truncation) or low similarity between the members of an *Alu* pair. The spacer (c) consists of non-*Alu* sequence between the aligned regions and represents the minimal estimated distance between the two *Alu*s. An internal flanking region is referred to as (b) so that the actual distance between aligned regions is the spacer (c) plus the internal flanking region (b) of an *Alu*. Therefore, the distance (b + c) is the actual distance between aligned regions of an adjacent pair of *Alu*s.

The information about *Alu* distributions is available at the Web site, and the extensive *Alu* pair tables (CC.html, CD.html, DC.html, DD.html) are interfaced to our database. The analysis utilized human sequence information available up to September 9, 1999. (Our system provides for the efficient incorporation of new data as they are added to the human genome database.) Included in the Web site is a description of the associated chromosome region or gene to aid in identifying regions that may be "at-risk" for possible *Alu*-associated changes (an example is presented in Fig. 2).

Preference in the Alignment Length Distribution of Paired *Alu*s

The adjacent *Alu*s first were classified according to relative orientation. Two-thirds of the *Alu* pairs (46,087 / 70,324) had both *Alu*s pointing in the same direction (Fig. 3). As expected, the direct repeats were equally distributed between the DD and CC orientations (23,318 and 22,769, respectively).

The pairs of *Alu*s then were analyzed according to their alignment lengths, which would correspond to the length of adjacent *Alu*s over which interactions might be possible. These distributions were expected to reflect *rules* about insertion and stability of *Alu*s. Pairs of *Alu*s were divided into five categories according to the alignment length '*a*' (in bp): >276, 201–275, 126–200, 51–125, and <50. As shown in Figure 3, nearly 75% of the *Alu* pairs contained aligned regions corresponding to nearly full-length (>276 bp) or half-length *Alu*s regardless of orientation.

Among all the *Alu* pairs, 41% of direct and 50% of inverted *Alu* pairs had alignment lengths >276 bps. The second most-frequent category of *Alu* pairs, containing 24% (17,007) of the total, was that for which the alignment lengths were approximately one half the length of an *Alu* consensus sequence (126–200 bp). This category may reflect the dimeric nature of the ancestral full-length *Alu* that is 282 bps long (excluding the poly(A) tails). The upstream half of an *Alu* sequence, which is derived from the free left *Alu* monomer (FLAM for review, see Jurka 1995), is 31 nucleotides shorter than the downstream region, which is derived from the free right *Alu* monomer (FRAM) that contains the polyA tail. In the ancestral *Alu*, the FLAM and FRAM, are joined by a poly(A) stretch, with the FLAM *Alu* monomer containing the split RNA polymerase III promoter.

Bias and Exclusion of Adjacent *Alu*s in Relation to Separation of Repeats

Previously, it was shown that distance was a major component in the distribution of *Alu* pairs, where the distance between *Alu*s

Alu Alignment Data

link to Legend

Locus ID#	Alu1 coordinates				Alu2 coordinates				num. of matches	num mis matches	simil arity score	percent identity	(a) align len.	(b+c) len.	(b) flank len.	(c) spacer len.	ttl. gap len.	num. of gaps	Loci Description	family1	family2
	fragment		aligned		fragment		aligned														
	start	finish	start	finish	start	finish	start	finish													
AC005500_13	86832	87131	86832	87130	87149	87468	87149	87447	245	49	0.82	82	299	19	1	18	5	5	Chromosome 22q11 PAC Clone p52f6 In DGCR Region, complete sequence.	Alu Jb	Alu Sz
X64467	9460	9753	9460	9753	9771	10070	9772	10065	240	51	0.81	82	294	19	1	18	3	5	ALAD gene for porphobilinogen synthase.	Alu Jo	Alu Jb

Figure 2 An example of an on-line summary of an adjacent pair of *Alus*. This table corresponds to a "cell" in the CD data table found at http://dir.niehs.nih.gov/ALU/CD.html. It shows the only two *Alu* pairs identified that meet the following criteria: the shared alignment length is between 200 and 276 bp, the paired *Alu*s are in the CD orientation, the separation distance b + c is between 21 and 40 bp, and there is between 86% and 90% identity. The Locus ID no., fragment coordinates, and family are extracted from the original map file. The number of matches, mismatches, similarity score, and the number of gaps are extracted from the alignment output. The locus description is extracted from the original GenBank sequence data file, as is the information necessary to determine the aligned sequence coordinates for the two fragments. Other features are described in greater detail in Methods.

was simply identified as non-*Alu* sequence between the pair of *Alu*s (Jurka 1995). Our analyses, based on the separation of aligned regions, are consistent with these findings. Furthermore, as discussed below, they suggest that alignment is an important component in the distribution of pairs.

As shown in Figure 4A, nearly 30% of all the direct *Alu* repeat pairs are separated by <20 bp, similar to the previous report where much less of the genome had been analyzed (Jurka 1995). There was little, if any, bias in the distances separating inverted repeats when only the distance between *Alu* sequences was considered, although the "tails-in" *Alu* pair category (DC) was greater for "c" <20 bp. For separations <20 bp, the total direct repeat pairs were ~10-fold more frequent than the total inverted repeats.

Analysis of *Alu* pair distribution on the basis of distance between aligned regions (b + c), rather than distance between *Alu* sequences revealed a much greater difference in frequencies between direct and inverted *Alu* repeats. Unlike for direct repeats, only a relatively small number of all the inverted *Alu* repeats were identified that had aligned regions separated by distances <20 bp (Fig. 4B). This may suggest that inverted *Alu* repeats are excluded during insertion and/or they are an unstable configuration. The total number of direct repeats separated by <20 bp was ~12- and 25-fold, more than the DC or CD classes of inverted repeats, respectively. For separations >20 bp, the total number of inverted and direct repeats was nearly constant for different size classes and within a factor of two of each other, suggesting more random insertion. It is

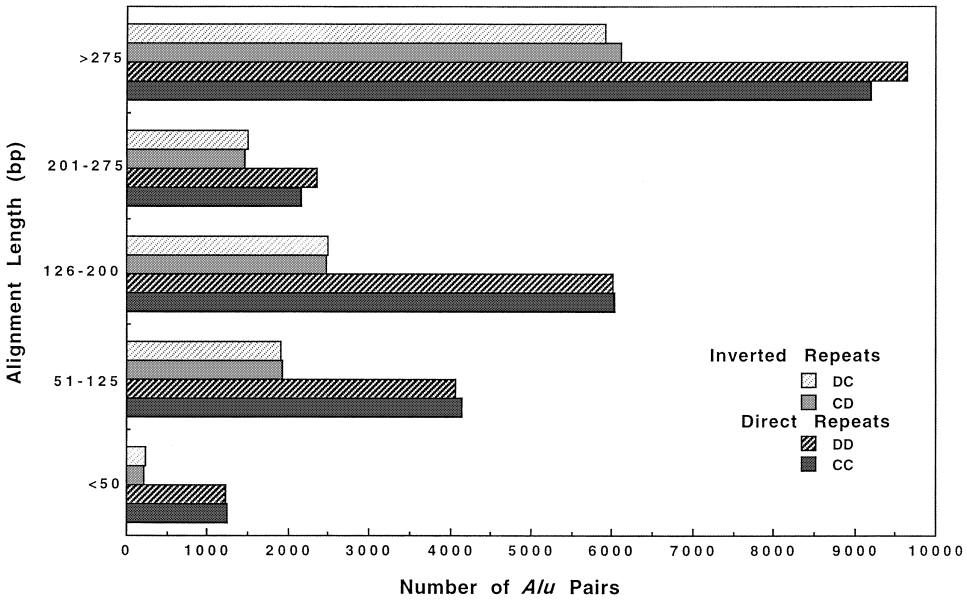

Figure 3 Length of alignment between adjacent *Alu* pairs. Pairs of *Alu*s were analyzed according to orientation and length of alignment. As noted in the figure, the *Alu* pairs were grouped into five groups of alignment lengths. Pairs separated by more than 650 bp were excluded from the analysis.

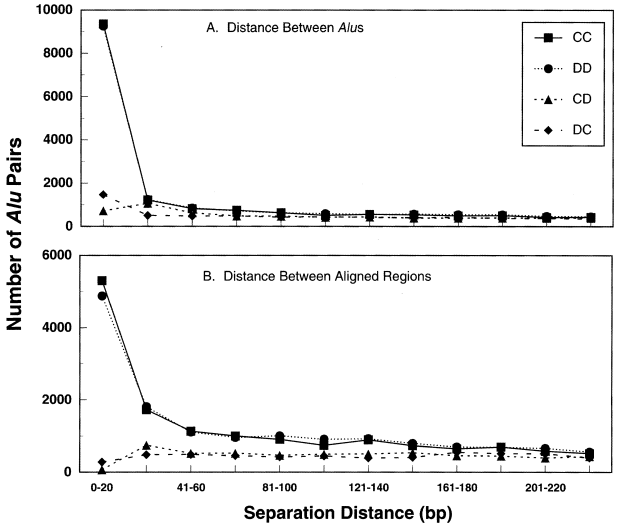

Figure 4 Distance between adjacent *Alu*s versus distance between aligned regions of adjacent *Alu*s. (*A*) Analysis of separation distance according to distance between *Alu* sequences. Nearly 30% of all direct adjacent *Alu*s are separated by only ≤20 bp. Beyond this distance, and also for all the inverted *Alu*s, there appears to be a random distribution of separation distances (that is, there is no preference in the distance separating adjacent *Alu*s.).(*B*) Analysis of separation distance according to nucleotides separating aligned regions. Inverted repeats, especially the "tails-out" (CD) category, were uncommon. Beyond 40 bp, there appears to be a random distribution of separation distances for both the inverted and direct *Alu* pairs. To determine the number of occurrences of *Alu* pairs separated by the unique spacer distance "c", we used the Perl program `Bins C`. For each of the four different orientation data sets, we ran `Bins` using 12 different window settings between 0 and 240 bp, divided into 20-bp bins. A similar program, `Bins 2`, was used to subdivide the data into groups depending on the "b + c" distance, representing the distance between aligned sequences (see Methods).

interesting that while the most frequent group of inverted *Alu*s are those separated by <20 bp (Fig. 4A), the actual distance between aligned regions often is 20 to 40 bp (compare Fig. 4A to 4B), suggesting that one member of the *Alu* pair is truncated (discussed below).

Given the importance of alignment in *Alu* pair distribution, the *Alu* pairs were analyzed according to the length of aligned region and either distance between *Alu* sequences (Fig. 5A-E) or distance between aligned regions (Fig. 6A-E). As shown in Figure 5A-E, large differences in the relative number of direct versus inverted repeats were found when the size of aligned regions was considered. For *Alu* pairs with approximately full-length alignments (>275 bp, the upper limit was arbitrarily chosen to be 500 bp, although few *Alu*s exceeded 300 bp) and half-length alignments (125–200 bp), there is a vast excess of direct versus inverted *Alu* pairs for short spacer distances (<20 bp). Surprisingly, there are many fewer direct repeats in the 200–275 bp and the <125 bp alignment categories. Possibly this pattern is a reflection of the dimeric nature of *Alu*s.

Regardless of the size of the aligned regions, there is clearly a reduction in inverted *Alu* pairs with short spacer distances (<20 bp) between the aligned regions (Fig. 6A-E), although the exclusion seemed somewhat less for the DC category of the approximately full-length *Alu*s. This is consistent with our previous results, where we found that for full-length *Alu*s, there was a strong bias against inverted repeats that are closely-spaced (Lobachev et al. 2000). There were more than twice as many CD-oriented pairs spaced between 21–40 bp apart as there were of DC pairs (459 versus 220, respectively). Interestingly, this difference applies only to the *Alu* pairs in the full-length alignment category (>275 bp)(discussed below).

Sequence Identity between Paired *Alu* Elements

An analysis of *Alu* pair distribution based on identity between the *Alu*s may reflect a role for homologous interactions in their appearance and stability. Most *Alu* pairs share between 65% and 85% sequence identity for both direct (Fig. 7A) and inverted (Fig. 7B) repeats. There were no apparent differences between distributions for the CC vs DD orientations or the CD vs DC inverted repeat orientations (data not shown). The dis-

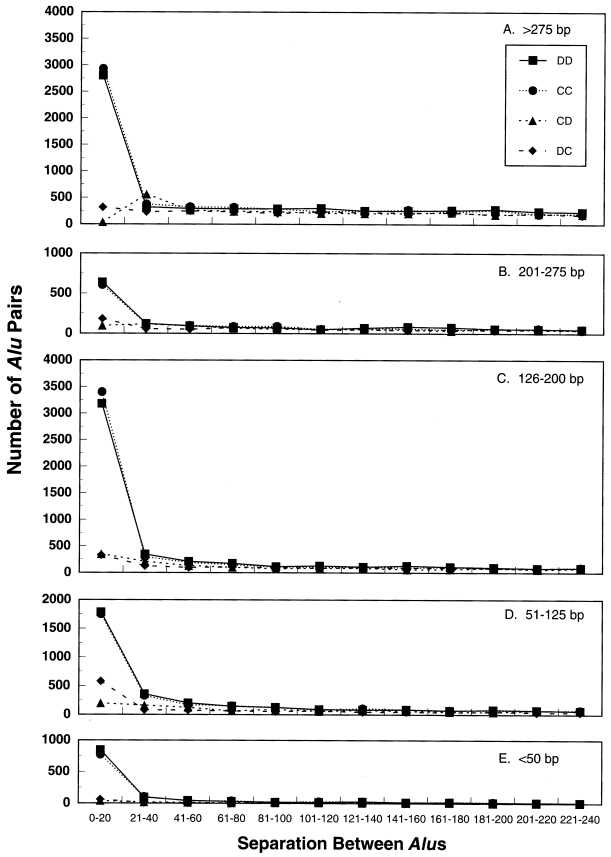

Figure 5 Distance between adjacent *Alu* sequences for *Alu* pairs having different alignment lengths. The adjacent *Alu*s were analyzed as described in Figure 3A, and the pairs were grouped into five alignment length ranges (bp): (*A*) >275 ; (*B*) 201–275; (*C*) 126–200; (*D*) 51–125; (*E*) <50.

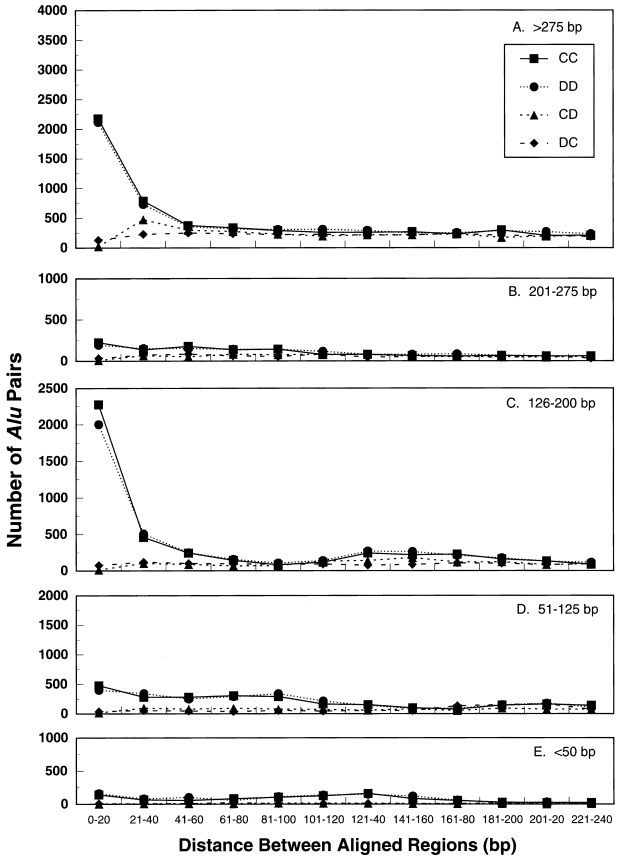

Figure 6 The distance between homologous regions for *Alu* pairs of different alignment lengths. The adjacent *Alu*s were analyzed as described in Figure 3B, where the separation distance "b + c" (see Fig. 1) and the pairs were grouped info five alignment length ranges (bp): (*A*) >275; (*B*) 201–275; (*C*) 126–200; (*D*) 51–125; (*E*) <50.

tributions were similar for alignment lengths up to 275 bp. The exception was for *Alu* pairs that had a short alignment length (<50 bp) where identities greater than 90% range were observed frequently. For the other categories (>50 bp), there were few *Alu* pairs that shared >90% identity. The degree of identity distribution for the approximately full-length direct and inverted *Alu* pairs were somewhat narrower and shifted towards greater identity, possibly suggesting interactions or conservation of some features of the *Alu*s. It will be interesting to determine if there are regions within the full-length *Alu*s that are more conserved.

Because the distributions of direct and inverted full-length *Alu* pairs differed dramatically for short separation distances between aligned regions (Fig. 6A), we examined further their distribution in order to evaluate the relationship between separation of aligned regions and level of homology. Presented in Table 1 are the frequency distributions for *Alu* pairs with long alignment regions (>275 bp) classified by levels of identity and separation distance between the aligned regions. Ninety percent of these *Alu* pairs exhibit 70%–90% sequence identity, with the number of pairs in the 70%–80% identity group about 1.7-fold greater than in the 80%–90% group, regardless of orientation within the pair (see Table 1). The aligned regions in one third of the direct *Alu* pairs are separated by 40 bp or less (Fig. 6 and Table 1).

While closely spaced, inverted *Alu* pairs generally are less frequent than direct pairs, our analysis revealed features in the distribution of full-length inverted *Alu* pairs that correlate with the degree of identity and orientation (i.e., tails-out or tails-in). As shown in Table 1, the proportion of inverted *Alu* pairs separated by (<20 bp with 80%–90% identity was considerably lower than for direct repeats, regardless of orientation of the inverted *Alu*s: ~250- and ~20-fold lower for the *Alu* pairs with tails external (CD) and tails internal (DC), respectively. Unlike the direct *Alu* pairs, the frequency appeared to increase somewhat with decreasing degree of homology. (Although the total number of events is small, these results are consistent with the observation of only one pair of inverted *Alu*s in the 90%–100% identity category up to 40 bp separation as compared to 14 among 34 direct *Alu* pairs.) The frequencies of CD and DC pairs were comparable at longer separation distances (41 to >80 bp) and were independent of homology; the frequencies also were more comparable to those of direct repeats. There was a difference between the frequencies of CD and DC repeats among the *Alu*s separated by 21–40 bp, with the CD pairs being about

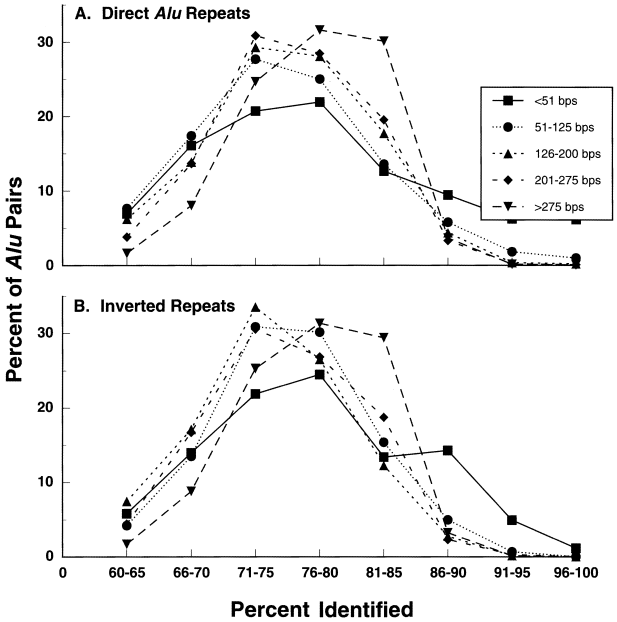

Figure 7 The frequency of *Alu* pairs with different levels of homology in their aligned regions. Presented is the distribution of direct (*A*) and inverted (*B*) *Alu* pairs sharing different lengths of alignment in relation to the extent of homology. The data are derived from all the *Alu* pairs shown in Figure 1.

Table 1. Distribution of *Alu* Pairs with Long Alignment Regions (>275 bp) According to Orientation, Separation Distance, and Sequence Identity

Orientation of paired *Alu*s	Distance between aligned regions (b + c) (bp)	Frequency of *Alu* pairs: degree of identity between paired *Alu*s 91%–100% (No.)	81%–90% (%)	71%–80% (%)	60%–70% (%)
Inverted					
DC	0–20	(0)	1.0	2.6	3.6
Total: 5786	21–40	(1)	3.0	4.2	4.3
	41–60	(0)	3.3	4.6	5.0
	61–80	(0)	3.6	4.0	5.1
	81–100	(2)	2.7	4.3	4.8
	>100	(6)	86.3	80.4	77.3
	(Total)	(9)	(1916)	(3255)	(606)
CD	0–20	(0)	0.1	0.5	0.5
Total: 5948	21–40	(0)	7.6	7.5	9.3
	41–60	(2)	4.4	4.8	6.8
	61–80	(0)	3.5	5.2	5.1
	81–100	(1)	3.8	3.5	5.0
	>100	(4)	80.6	78.5	73.5
	(Total)	(7)	(1996)	(3354)	(591)
Direct					
(DD + CC)	0–20	(11)	24.3	22.2	20.9
Total: 18,253	21–40	(3)	7.4	8.3	9.3
	41–60	(2)	3.9	3.8	4.7
	61–80	(2)	3.3	3.6	3.4
	81–100	(1)	2.9	3.2	4.0
	>100	(18)	58.3	58.9	57.8
	(Total)	(37)	(6205)	(10,251)	(1760)

twice the frequency of the DC pairs and exhibiting frequencies in the range of the direct repeats. As presented in the Discussion, the rarity of inverted *Alu* pairs that have closely spaced, highly related regions (e.g., <20 bp), regardless of heads-in or heads-out orientation, is likely because of a destabilizing effect resulting from an interaction between the homologous regions.

Truncations of Large *Alu*s

The similar frequencies for inverted pairs (especially CD) and direct repeat pairs at distances >20 bp may indicate that targeting mechanisms exist for inverted repeats (or at least for the CD category) as well as direct repeats (Jurka 1997). However, at short distances, the high insertion rate may be counteracted by the instability of inverted repeats. The differences in distribution between the direct repeats and inverted repeats that are CD or DC led us to analyze *Alu*s that are truncated relative to each other within a pair with a view to understanding how they may have originated. For example, the spacer length between aligned regions appears to reflect the insertion and/or stability of repeats. The spacer between aligned regions of adjacent *Alu*s could be formed by non-*Alu* sequence or derived, in part, from asymmetric truncations of adjacent *Alu*s. We, therefore, examined whether adjacent *Alu* pairs were truncated by more than 5 bp relative to each other at the 5′ ends (heads). (The poly(A) 3′ tails were not analyzed because they arise through expansion and their lack of uniformity made such an analysis technically unfeasable.) The truncations were identified among the pairs of *Alu*s with >275 bp alignment each and these were categorized according to whether the 5′ truncated ends were internal (CD category) or external (DC category) (Table 2).

For the heads-out repeats, approximately one half (44%–55%) of the *Alu* pairs have a truncated *Alu*, regardless of the distance between the *Alu*s. This contrasts with the heads-in inverted repeats and the direct repeat categories, which also were markedly different from each other. For the direct repeats, 45% of the pairs have a truncation when the *Alu*s are closely spaced (<20 bp). However, the truncations are not evenly distributed: there is a strong bias (10:1) for truncations of the 5′ ends of *Alu*s that are internal to the closely spaced pairs. At greater distances, the frequency of truncations increases and the bias disappears, with the ratio of internal truncations becoming somewhat less than external. For the heads-in category (CD), there is a strong bias towards equal alignment lengths (no truncation) when the spacer is short (20–60 bp). At longer distances, the frequency of truncations is comparable to that for the direct and the heads-out inverted repeat categories. Among the few heads-in *Alu* repeats that are found at very short distances (<20 bp),

Table 2. Truncation[a] of "Heads" in *Alu* Pairs According to Orientation and Distance

Orientation of pairs	Separation distance (b + c) (bp)	Frequency			
		internal truncation %	external truncation %	no truncation[b] %	total no.
Direct Repeats					
CC	0–20	39.8	4.2	55.9	10170
	21–40	36.3	17.8	45.9	3541
	41–60	34.4	26.8	38.8	2242
	61–80	27.0	31.5	41.5	1962
	81–100	24.4	39.2	36.3	1923
	101–120	24.8	35.3	40.0	1655
Inverted Repeats					
CD	0–20	50.0	N/A	50.0	66
	21–40	13.8	N/A	86.2	747
	41–60	28.4	N/A	71.6	524
	61–80	36.2	N/A	63.8	525
	81–100	38.3	N/A	61.7	468
	101–120	50.8	N/A	49.2	496
DC	0–20	N/A	56.4	43.6	280
	21–40	N/A	48.9	51.1	485
	41–60	N/A	47.4	52.6	490
	61–80	N/A	46.4	53.6	455
	81–100	N/A	45.2	54.8	423
	101–120	N/A	48.6	51.4	449

[a]Truncations are ≥5 bp
[b]Includes truncation ≤5 bp

they are nearly equally distributed between truncated and equal sizes. We suggest that these differences in truncation patterns reflect differences in mechanisms of insertion and/or stability (see Discussion).

Age of *Alu*s in Pairs Differs with Distance

Recently retroposed *Alu* sequences usually are not fixed in the population (Batzer et al. 1996) and are, therefore, presumed to be a source of ongoing genomic diversity. Amplification and mobilization is limited to the young *Alu* subfamily sequences (i.e., *Alu*-Y, *Alu*-Ya1, *Alu*-Yb8, *Alu*-Ya5, and *Alu*-Ya8) that have been inserted in the last 1.5 million years (Arcot et al. 1995). We therefore analyzed the distribution of *Alu* pairs in order to explore the relationship between age, orientation, and distance between aligned sequences. The analysis, which was limited to approximately full-length sequences that were at least 80% identical, compared the incidence of closely spaced *Alu* pairs (b + c <40bp) and more distantly spaced *Alu*s (b + c >41 bp) containing at least one member of the young *Alu* subfamily (Table 3). Young *Alu*s were nearly 30% more frequent among the closely spaced pairs (<40 bp), regardless of orientation. While the reasons for this preference are not apparent, it is possible that the modest bias for young *Alu*s reflects a more recently evolved targeting mechanism or a slow mechanism for separating them.

DISCUSSION

General Approach to Investigating *Alu* Distribution

Understanding the organization of *Alu*s in the human genome is expected to shed light on *Alu* integration, *Alu* changes, and the potential for *Alu*s to affect genome stability. Our approach incorporated newly developed computational tools along with previously developed programs to analyze *Alu* pairs in terms of the potential for homologous interactions.

The pairwise approach to analyzing *Alu*s was motivated in part by observations from several model systems. Pairs of large inverted DNA repeats can be unstable, lead to deletions, and stimulate recombination between DNAs surrounding the inverted repeats in yeast (Gordenin et al. 1993; Ruskin and Fink 1993; Nag and Kurst 1997; Tran et al. 1997; Lobachev et al. 1998, 2000;) and mammalian cells (Akgun et al. 1997; Lewis 1999; Lewis et al. 1999). Sequence identity was suggested as a likely factor in the stability of *Alu* repeats based on results in yeast. Lobachev et al. (1998) demonstrated that large palindromes could stimulate recombination nearly 10,000-fold in yeast. Recently, we established that the shorter inverted human *Alu* repeats, but not direct *Alu*s, also could stimulate recombination nearly 2000-fold in yeast between surrounding DNA and an ectopic allele when the *Alu*s were separated by <20 bp (Lobachev et al. 2000). The recombination rates decreased exponentially as the

Table 3. Influence of Orientation and Separation Distance on *Alu* Pairs Containing a Young Subfamily Member

			Young *Alus*[b]	
Orientation	**Separation distance[a] (b + c) (bp)**	**Sample size**	**no.**	**%**
Inverted				
CD	0–40	815	175	21.5
	41–80	1049	187	17.8
	>80	8082	1319	16.3
DC	0–40	765	173	22.6
	41–80	945	174	18.4
	>80	7907	1396	17.6
Direct				
CC	0–40	7031	1580	22.5
	41–80	2133	353	16.5
	>80	10890	1810	16.6
DD	0–40	6689	1493	22.3
	41–80	2071	318	15.3
	>80	11707	1977	16.8

[a]b + c (in bp)
[b]*Alu* subfamilies *Alu*-Y, *Alu*-Ya1, *Alu*-Yb8, *Alu*-Ya5 and *Alu*-Ya8

distance was increased to 100 bp. The recombination also was reduced as sequence identity decreased, with 80% appearing to be the lower limit for stimulation of recombination by inverted *Alu* repeats. Mutants were identified in which diverged and distantly separated (100 bp) inverted *Alu* pairs also were capable of stimulating recombination, suggesting that these inverted repeats are potential ARMs. These observations along with previous computational studies on the distribution of *Alu* pairs (Jurka 1995), strongly supported new approaches to an investigation into the pairwise distribution of *Alu*s and provided insight into what parameters should be considered.

Our study focused on four attributes of *Alu* pairs that might be important in the integration and stability of *Alu* sequence pairs in genomes of humans: (1) Orientation of each member of the pair, (2) size of the inverted repeat, (3) distance between the aligned regions of the pair, and (4) sequence identity. In addition, we examined the age of *Alu*s. The present study considerably extends our recent report demonstrating a reduction in closely spaced inverted *Alu*s as compared to direct *Alu* repeats (Lobachev et al. 2000). Based on results from model systems, other factors also are likely to be important in the stability of *Alu* sites including genetic background (Gordenin et al. 1993; Ruskin and Fink 1993; Tran et al. 1997; Lobachev et al. 1998, 2000;). For example, recombination in mammalian cell constructs containing direct identical *Alu* repeats is high in p53 defective as compared to p53 wild-type cells (Gebow et al. 2000). We suggest that nucleotide sequence, mutagenic agents, and DNA metabolic processes, as well as variation in levels of relevant proteins, may contribute to the ability of inverted *Alu*s to stimulate genomic change.

Direct *Alu* Repeats and Preferences

It is clear that orientation and distance between aligned regions are important factors in *Alu* distribution. For long separations (i.e., >80 bp) between aligned regions, there do not appear to be any preferences or exclusions of direct or inverted *Alu* pairs.

In this study, we confirm that there is a vast excess of closely spaced *Alu*s in the direct as compared to the inverted orientation (Jurka 1995). This excess has been proposed to result from targeted *Alu* integration to an approximate 15 bp sequence (Jurka 1997). The sequence was identified by the high degree of conservation at the ends of the target represented by 5′ AAAA and TYTN 3′ consensus sequences. In addition, the target is preceded by a 5′ TT, producing the strong 5′ TT/AAAA consensus sequence where "/" indicates the position of a proposed nucleolytic cleavage (nicking) that would occur on the opposite strand (Fig. 4 in Jurka 1997). A second nicking site has been proposed to occur immediately after the 5′ TYTN/consensus sequence and marks the integration position of the incoming *Alu*. The final arrangement appears as follows: 5′-TTAAAAnnnnnnnnnnTYTN(*Alu*) where "n" stands for any base; the number of the "n" nucleotides in the middle of the target can vary. (The conserved consensus sequences are indicated by upper-case letters, including the 5′ TT immediately preceding the integration target.) Hexamers derived from the TTAAAA consensus such as TTAAGA, TTAGAA, CTAAAA, TCAAAA etc., also are associated with targets for *Alu* integration (Table 3; Jurka 1997). Therefore, the likelihood of integration is expected to increase with abundance of the target hexamer sequences. Because *Alu* sequences frequently contain TTAAAA-like signals at the 3′ ends, the tails of *Alu*s are good targets for subsequent *Alu* inte-

grations. This leads to the prediction that new *Alu*s will integrate about 15 bp from the nicking site "/" within the TTAAAA-like hexamer, and in the same orientation as the preexisting *Alu*.

It should be noted that complementary AATTTT-like signals within preexisting *Alu*s would determine integration of the incoming *Alu*s in the opposite orientation. However, in directly oriented *Alu*s, the complementary AATTTT-like target signals are about 10 times less frequent than the TTAAAA-like signals (Jurka, unpubl.). Therefore, closely spaced *Alu* integrations in direct orientation (i.e., CC or DD) would be expected to be around 10 times more frequent than inverted sequences (CD or DC). (Furthermore, the AATTTT-like signals are scattered randomly in *Alu*s so that no systematic pattern of CD or DC pairs would be expected.) Regardless of targeting preferences, this would not explain the rarity of closely spaced inverted *Alu*s (<20 bp) as compared to *Alu*s that are more distant.

The present study differs from previous approaches in that we have classified *Alu* pairs according to the distance between aligned regions rather than just the distance between *Alu* sequences. Because the actual distance between aligned regions (b + c) is greater than or equal to the distance between *Alu* sequences (c), it appears that many of the closely spaced pairs may contain truncated *Alu*s. Our extensive analysis of approximately full-length *Alu*s (Table 2) has demonstrated that most of the truncations are in the internal head of the direct repeats. It is interesting that the strong preference for closely spaced aligned regions of *Alu*s only applied to the largest categories of direct repeat pairs, >275 bp and 125–200 bp. Possibly this is a result of an integration preference for full-length and half-length *Alu*s (i.e., one member of the dimer within an *Alu*).

Both the direct and the inverted *Alu* pairs that have full-length alignments tend to be more closely related than when the alignments are shorter. Among the reasons are that some regions diverge less readily than others, the full-length *Alu*s have long polyA tails that would contribute to overall sequence identity, and possibly there are greater opportunities for homologous interactions. Also, long *Alu*s tend to be younger and less diverged (Arcot et al. 1995; Batzer et al. 1996).

Inverted *Alu* Pairs and Exclusion

We found that, unlike for direct repeats, inverted *Alu* pairs with closely spaced (<20 bp) aligned regions were uncommon regardless of size of alignment or orientation (Fig. 4B) and especially rare among the *Alu* pairs with nearly full-length alignments. As the distance between aligned regions increased beyond 20 bp, the frequency of direct and inverted *Alu* pairs became more uniform, suggesting random integration. Interestingly, for *Alu*s whose alignment regions are separated by <20 bp nucleotides, the CD pairs (tails out) are even more excluded than the DC pairs. Possibly, this is a result of more heterogeneity of tails versus the unique sequence of *Alu* heads.

The exclusion of closely spaced inverted repeats is consistent with the observation that inverted repeats at close distances are unstable in yeast and, for the case of *Alu*s, the instability is highly dependent on distance (Lobachev et al. 1998; Lobachev et al. 2000). We propose that the observed absence in the human genome of closely spaced inverted *Alu*s that are highly related is a result of this motif adopting a noncanonical DNA form, including intrastrand pairing, that has a high potential for stimulating chromosomal changes. Based on results in yeast, closely spaced *Alu*s may be sites of genetic instability in normal cells or in cells that are defective in repair or some aspect of DNA metabolism (Lobachev et al. 2000). Several human diseases have been found to be associated with changes in regions of inverted *Alu*s (Deininger and Batzer 1999). We also suggest that genetic factors similar to those affecting stability in yeast may extend to humans. While mutations in mismatch repair did not affect the instability of diverged *Alu*s in yeast (Lobachev et al. 2000), it remains to be determined if systems responsible for mismatch recognition may help to prevent *Alu* associated instability in human cells.

Thus, we conclude that sequence identity and distance are important factors that contribute to the distribution of *Alu* pairs and the potential for *Alu* pairs to cause genome stability. These observations with inverted as well as direct repeats may be useful in understanding the frequent clustering of *Alu*s.

Truncations of Closely Spaced *Alu* Pairs and Mechanisms of Integration and/or Instability

We found that for heads-out repeats (DC) and distantly spaced heads-in (CD) and direct repeats, about one half of the pairs had a truncation in one of the *Alu*s. Departures from this frequency may indicate factors that affect *Alu* insertion and/or stability of *Alu* pairs. For closely spaced direct *Alu*s, the internal *Alu* head frequently is truncated relative to the external head. The results with the inverted *Alu*s clearly are different. The closely spaced (20–60 bp) heads-in inverted *Alu* pairs have a strong bias towards *Alu*s with no 5′ truncations, while there is no such bias for the heads-out category. (The higher frequency of truncated pairs among the very few with aligned regions that are <20 bp apart may simply reflect the instability of inverted *Alu*s in close proximity.)

The differences between the two categories of inverted repeats and the markedly different observations with the direct repeats have interesting implications both for the origin of *Alu* repeats and their potential instability. Models based on a simple retroposition at

the site of integration do not account for the various truncation patterns we observe: Fifty-percent truncation of one of the *Alu*s in distantly separated pairs of *Alu*s, preferred truncation of the internal head of direct repeats, and no truncation of heads-in inverted repeats. As discussed above, there is a strong preference for insertion of an *Alu* in a direct orientation immediately next to the tail of an existing *Alu*. The present results suggest that along with targeting, there is an associated removal of some of the 5′ end of the incoming *Alu* and as targeting next to an *Alu* become less likely (i.e., for more distantly spaced repeats), the truncations are less frequent.

The reasons for the lack of truncation in the heads of the closely spaced heads-in *Alu*s but not heads-out inverted *Alu* pairs are not clear. However, they may indicate a directional complementary pairing mechanism prior to integration that starts from the apex of the inverted pair (i.e., the closest sequences in an *Alu* pair). This would result in a preference for full-length pairing at the apex that could be detected in the heads-in category but not the tails-in category (variability in size of poly AT tails would preclude such analysis of the tails-in pairs). Opportunities for complementary interactions might arise if *Alu* RNA had been reverse transcribed to cDNA prior to integration. Although there are few direct examples, genetic evidence from yeast are consistent with cDNAs being an intermediate in recombination (Derr and Strathern 1993). More directly, RNA injected into mouse (Giordano et al. 2000) has been shown to yield cDNA. If the cDNA interacted with complementary displaced single-strand flaps that arise during normal DNA replication, this would result in stabilized double-stranded flaps that would be resistant to cleavage by the Fen1 flap-processing enzyme and a stabilized inverted repeat in the lagging strand (Gordenin et al. 1997).

Another possible explanation has to do with the transcription by RNA polymerase III. In the DC class of inverted repeats, 5′ to 3′ transcription from both ends converges toward the center of the repeat. The transcription diverges outward from the center in the CD class. DC repeats therefore can be opened by transcription for recombination because transcription initiating in D's promoter could proceed through the center and into the internal middle run of Ts between the FLAM and the FRAM in the second *Alu* in the pair. This would be precluded if the C in the DC pair presents a "poly T" tail so that transcription is halted before the second *Alu* creating the possibility of ATA triplexes from the interaction between single stranded transcribed poly A and the AT duplex in the DC pair.

Genes and Regions with *Alu* ARMs in the Human Genome

This study was motivated in part by the association of numerous human disorders with *Alu*-mediated sequence rearrangements. However, given the high frequency of *Alu* sequences in the human genome, the number of *Alu*-associated diseases would appear low. The majority of the *Alu* sequences are stable, having remained in their present position in the human genome for millions of years and may pose relatively little threat to human health. However, we reasoned that a study of the characteristics of *Alu* pairs might reveal situations that potentially are unstable. These could include closely spaced, inverted *Alu* pairs of pairs that might be unstable in some backgrounds [as found for yeast (Lobachev et al. 2000)] or in response to environmental challenges.

Because we found that both distance and sequence identity were important factors in defining the distribution of *Alu* pair repeats and that they influenced the stability of inverted repeats in model systems, these factors may be useful in predicting regions or genes in the human genome that may be at-risk for instability as a result of inverted *Alu* pairs, based on information available from yeast and the similarity between systems in yeast and humans that deal with genetic stability (Resnick and and Cox 2000). Genes were, therefore, identified that might be at-risk for instability using criteria developed in yeast where 80% identity was the apparent minimum level of detectable recombinational impact of inverted *Alu*s in wild-type strains and inverted highly related *Alu*s separated by 40 bp could induce a nearly 10-fold increase in recombination (see Lobachev et al. 2000). We consider these as conservative parameters for identifying at-risk regions because changes in genetic background can turn genetically stable, inverted *Alu* repeats into stimulators of recombination (Lobachev et al. 2000). Presented in Table 4 are loci that were found to contain inverted *Alu* pairs in which the separations between aligned regions were <40 bp and there was at least 80% identity between the aligned regions. Many *Alu* pairs (268) also were identified whose loci have not been ascribed a function and many of which share >86% sequence identity (data not shown).

Loss of heterozygosity (LOH) of several genes on the list have been linked to various genetic illnesses. These include *TNX*, encoding tenascin-x (associated with an undesignated Ehlers-Danlos syndrome type, clinically similar to type II, but with distinct ultrastructural characteristics) (Burch et al. 1997), *MeCP2*, encoding x-linked methyl-CpG-binding protein (Rett syndrome) (Amir et al. 1999) and *GPC3,* associated with Simpson-Golabi-Behmel syndrome, among others. Four of the gene-associated regions contained more than one set of the *Alu* inverted repeats: the T-cell receptor α δ locus, the LIM kinase1 (*LIMK1*) gene, the *GART*- and *AML*-related genomic DNA, and the basic TFII p44 subunit gene. Because LOH can occur through recombina-

tion, it is possible that LOH might be the result of recombination that is stimulated by inverted repeats.

A deletion of the *LIMK1* gene coding for the neuregulin-interacting serine, threonine, tyrosine kinase, is associated with the less-severe phenotype of Williams-Beuren syndrome (Robinson et al. 1996), characterized by diminished visuospatial construction cognition and low IQ (Frangiskakis et al. 1996). The more severe phenotype that also includes vascular disease is linked to a more extensive deletion on 7q11.23 that results in a mutated elastin gene (Frangiskakis et al. 1996). This gene may be prone to *Alu*-facilitated rearrangements. It is striking that the *GART*- and *AML*-related regions contain four closely related, large *Alu* pairs that are separated by less than 40 bp. Also noteworthy is the Huntington's disease region that, in addition to containing one *Alu* pair that met the above criterion, also had more *Alu* pairs that had similar features. This region also contains a multiple trinucleotide repeat at-risk DNA motif (Aronin et al. 1995), however, this does not exclude the possibility of another risk factor.

It is interesting that λ DNA immunoglobulin light chain DNA has many inverted *Alu* pairs, one of which satisfies the criteria of Table 4. Possibly, they provide another means for generating antibody diversity. Of particular interest is the breakpoint cluster region gene *BCR,* where *Alu*s within this region have been implicated in chromosomal translocation leading to Ph1 + *bcr*- acute leukemias (Chen et al. 1989). The identifica-

Table 4. Closely-Spaced (≥40 bp) and Related (>80% Homology) Inverted *Alu* Repeats Associated with Known Genes[a]

Locus ID[b]	Locus description[b]	Indent.[c] (%)	Len.[d] (bp)	Dist.[e] (bp)	Subfamily 1st	Subfamily 2nd
CD orientation						
D26607	Endothelial nitric oxide synthase gene	81	109	7	Sxz	Y
Y11950	*PHKG2* gene, exon 16	86	180	13	Y	Sc
X64467	*ALAD* gene for porphobilinogen synthase	82	294	19	Jo	Jb
AP000113	*GART* and *AML* related genomic DNA chr 21q21.1	83	301	25	Sq	Sp
AE000658	T-cell receptor α; and δ; locus	88	131	27	Sp	Sc
AE000658	T-cell receptor α; and δ; locus	83	165	28	Spqxz	Sg
L78810	ADP/ATP carrier protein (ANT 2) gene	83	296	28	Y	Sx
U62293	LIM kinase 1 (*LIMK1*)	86	264	30	Sz	Sc
AF042084	Heparan glucosaminyl N-deactylase/sulfotransferase	82	298	31	Sx	Sp
AJ000673	CD94 gene (NK cell receptor) exons 4, 5, and 6	87	311	31	Y	Y
D28126	ATP synthase α; subunit gene	82	298	31	Sp	Sx
M90058	Serglycin gene, exons 1, 2, and 3	82	301	31	Sp	Sg
U63721	Elastin (*ELN*) gene, partial cds, and LIMK1	84	263	31	Sz	Sc
AP000114	*GART* and *AML*-related genomic DNA chr 21q21.1	81	266	32	Sg	Sc
AF030876	MeCP2 locus, X chromosome	81	306	33	Sx	Sp
AF019413	HLA class II region containing tenascin x (TEN-X)	81	126	34	Sz	Jo
Z68226	Huntington's disease region, cosmid L141A8	86	306	35	Y	Ya1
L06849	CD36 (macrophage type B scavenger receptor) gene	84	296	35	Sp	Sq
AP000152	Down's Syndrome Critical Region, chr 21q21.2	85	311	36	Sx	Sx
M31651	Sex hormone binding globulin (*SHBG*) gene	81	299	37	Sz	Sp
U62293	LIM kinase 1 (*LIMK1*)	81	295	38	Y	Sg
D87024	λ DNA for immunoglobulin light chain	84	305	39	Ya1	Y
DC orientation						
AF060911	Epithelial sodium channel α; subunit, exon 3	83	297	12	Sx	Sp
AP000116	*GART* and *AML* related genomic DNA chr 21q21.1	82	287	15	Sq	Sx
AF003529	Glypican 3 (GPC3) gene	82	212	17	Y	Sg
U29953	Pigment epithelium derived factor gene	80	310	17	Y	Sz
U80017	Basic transcription factor 2 p44 (btf2p44) gene	83	166	19	Sx	Sp
U07000	Breakpoint cluster region (*BCR*) gene	80	296	20	Sp	Y
AP000119	*GART*- and *AML*-related genomic DNA chr 21q21.1	82	141	34	Sg	Sxzg
U80017	Basic transcription factor 2 p44 (*btf2p44*) gene	95	311	35	Y	Y
AF088219	Beta (CC) chemokine gene cluster	84	302	36	Sq	Y
Z73359	*BRCA2* gene region	81	313	37	Sx	Sz
M27148	Alpha 2 plasmin inhibitor allele B	81	194	40	Spqxz	Spqxz

[a]Clones without a known function are not included. List is sorted according to separation distance. Minimum cutoff values used were >100 bp for length and >80% sequence identity. The maximum separation distance is 40 bp.
[b]The Locus identification (Locus ID) number and description are taken from GenBank.
[c]The percent identity (Ident.) between the two paired *Alu* sequences = number of matched nucleotides/alignment length.
[d]The alignment length (Len.) = number of matches + number of mismatches + total gap lengths.
[e]The total separation distance (Dist.) = b + c (see Fig. 1).

tion of this gene in our analysis supports our approach to identifying potential *Alu* ARMs.

We also have identified many genetically uncharacterized regions that contain *Alu* pairs with closely spaced, aligned regions (data not shown but available in web site database). Even if these regions are not associated with genes, they might be at-risk if the *Alu* pairs could initiate genomic changes such as LOH, chromosome loss [as found for yeast (Lobachev, unpubl.)] and translocations. They may correspond to highly polymorphic sites between individuals. It would be interesting to follow the stability of the various closely spaced inverted *Alu* pairs under different growth and exposure conditions. Based on results with yeast, mutants defective in DNA metabolism or expressing altered DNA metabolic proteins also may reveal situations under which closely spaced, inverted *Alu*s would be unstable.

METHODS

Computational Approaches and *Alu* Identification

The computational methods used in this study are listed in Table 5. They are available at the website http://dir.niehs.nih.gov/ALU/methods.html, where they are described at length. Computations were performed on either a SUN Sparc station running SunOS 5.5 (SUN Microsystems) or a Silicon Graphics O2 workstation running IRIX 6.3.

Sequences that were related to a consensus *Alu* sequence were identified. A map file of all human *Alu* sequences as of September 1999 was developed by comparing human genomic sequences in the GenBank database (release112.0, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland) with a well-defined *Alu* consensus sequence (Jurka 1993). The map file contains the following information for the 153,645 *Alu* sequence elements identifed (estimated to be between 14% and 30% of the total *Alu* sequences present in the human genome) in columnar form: locus, beginning and ending sequence coordinates, family of *Alu* sequence, the orientation of the sequence (D, denoting direct and C, denoting complementary), the percentage identity to the *Alu* consensus sequence, the ratio of mismatches to matches, and the alignment score.

The annotated sequence files were extracted from the GenBank database to create a GenBank sublibrary of *Alu* sequences. This sublibrary was needed to generate alignments between the *Alu* pairs.

Categorization of Pairs According to Relative Orientation of *Alu*s

The revised map file was used as input to derive a list of loci and their corresponding coordinates for each pair of adjacent *Alu* sequence. The program PFOLLOWS3 (Klonowski and Jurka 1997), which locates adjacent pairs, handled this task by taking the input file, minimum acceptable distance (0 bp), and maximum acceptable distance (650 bp) as parameters. These initial distances represented the "c" distance, or the minimum distance between one *Alu* consensus sequence and the next (see Fig. 3). The PFOLLOWS3 output contained four lines, one for each of the four possible orientation permutations for a pair and each with several thousand characters. Direct repeats could be in the complementary-complementary (CC) orientation (3′ <− 5′ 3′ <− 5′) or the direct-direct (DD) orientation (5′ −> 3′ 5′ −< 3′). The inverted repeats could be CD-oriented inverted repeats, which are oriented such that the 5′ ends, as related to the direction of transcription of active *Alu*s, are towards each other (3′ <− 5′ 5′ −> 3′), or DC oriented inverted repeats (5′ −> 3′ 3′ <− 5′). All pairs in the CC (complementary-complementary) orientations are listed on the first line of the output, followed by all the pairs in the CD (complementary-direct) orientation, the DC orientation and the DD orientation.

Four simple programs, written in Perl, were run by an executable script to extract the coordinates for each of the four types of pairs (CC, CD, DC, DD) from the PFOLLOWS3 output file. The four subfiles were reformatted to place the pair coordinates side-by-side in list format for subsequent reference and analysis. These coordinate list files were then used to extract the sequences from the GenBank-derived sublibrary by the program VEXT (Klonowski and Jurka 1997).

Extraction and Alignment of Paired *Alu* Sequence Fragments

Sequences from the sublibrary were extracted for both *Alu* elements in each pair. For the inverted repeats (the CD and DC *Alu* pairs), the reverse complementary sequence of the second *Alu* in each pair was generated for alignment. The *Alu* pairs that were direct repeats (CC and DD) were aligned with each other; for the inverted repeats, the first *Alu* in each of the pairs was aligned with the complement of the second *Alu* element. These pairwise alignments yielded various alignment characteristics [such as mismatches and matches (Waterman 1984; Faulkner and Jurka 1988)] for subsequent analyses of the various pairs of adjacent *Alu*s.

Retrieval of Coordinates

Once the sequences were aligned, the actual aligned coordinate numbers (the alignment program renumbered the coordinates to start with one) were regenerated using the short program PRENUM02 (Klonowski 1998). Summary data were extracted through execution of the UNIX grep command twice, first to get the recovered alignment coordinates, and then to retrieve the alignment statistics that were on separate lines. These files were pasted so that all the pertinent information was on one line.

Alignment Length and Percent Identity within Pairs of *Alu*s

The data were reformatted and files containing chromosome location, *Alu* sequence, and pairwise alignment characteristics were merged (the loci names were matched to establish correct merging of files). This enabled the alignment length "a", percentage identity, and other parameters to be determined for pairs of *Alu*s. An example of the parameters associated with each *Alu* pair in the website is provided in Figure 2. The alignment length "a" ("align len." in Fig. 2) is obtained as follows: if the aligned sequence of *Alu*1 (a1) > the aligned sequence of *Alu*2 (a2), then a = a1, otherwise a = a2. (The actual length a1 = ("*Alu*1 aligned finish") − ("*Alu*1 aligned start") + 1, and a similar procedure is used for a2.) The alignment length "a" also is equivalent to the "num(ber) of matches" + the "num(ber) of mismatches" + the total ("ttl. ") gap length. Individual gap lengths correspond to the deletion

Table 5. Program Command Procedures to Interface with GenBank Data Base and Generate *Alu* Distribution Data Set[a]

Filename	Function	Invoke command
C programs		
pfollows3	Finds adjacent pairs in map file within a specified distance	pfollows3 map 0 650 > source
vsub2	Extracts library sequence files specified in map file	vsub2 map [library] Out
vflop[b]	Changes the position of the columns	vflop
vext[b]	Extracts sequences of given fragment boundaries from Out	vext
pplan[b]	Sequence annotation, makes loci names unique	pplan
pcomp1	Get the reverse complement of the sequence	pcomp1 cd.nd.uniq cd.nd.comp[c]
pflank3	Aligns the first Alu sequence in pair with the second	pflank3 [.uniq] [.comp] [.align]
prenum2	Gets original coords. from .unique files and restores to .align	prenum2 [.align][.uniq1] [.uniq2] [align2]
Perl programs		
get_cc_pairs	Extracts CC pairs from pfollows3-generated output	get_cc_pairs source > cc
get_cd_pairs	" CD " " " " "	get_cd_pairs source > cd
get_dc_pairs	" DC " " " " "	get_dc_pairs source > dc
get_dd_pairs	" DD " " " " "	get_dd_pairs source > dd
reformat_grep	Calculates the alignment length and percent identity	reformat_grep cd.grep > cd.grep1
coordinates	Gets the unaligned fragment coordinates from seq. file	coordinates [.uniq]
get_descrip	Gets the gene/cosmid sequence description " " "	get_descrip [.uniq]
Bins	Echoes the alignment stats for data within a length range	Bins [low] [high] [infile]
Bins2	Sorts data according to (b+c)	Bins2 [low] [high] [infile]
Bins3	Sorts data according to % identity	Bins3 [low] [high] [infile]
Shell scripts		
CON1	Translates multiple spaces into a single space	CON1 ../cd > cd[c]
CON	Translates spaces to new lines (tr " " "\n" <$1)	CON cd > temp[c]
t	Calls vflop and vext to extract sequences[d] from Out	t
rename	Sends parameters to pplan appends extension .uniq	rename cc.st[c,e]
ext_coord	Calls coordinates, vflop and get_definitions	ext_coord cc.nd.uniq[c,e]
Batch files		
Reformat	Invokes CON1, and CON and renames temp [input]	Reformat
Extract Pairs	Calls t for cc, cd, dc, and dd files	Extract Pairs
Make_uniq	Calls rename for cd.nd, cd.st, cd.tot etc.[c,e]	Make_uniq
Batch_align	Calls pcomp1 and pflank3 to align all .st w/ .nd files	Align_in_batch
Mv_align	Renames all .align2 files .align files	Mv_align
Grep_ac	Gets the coordinates for the aligned sequences	Grep_ac
Grep_as	Gets the alignment statistics from the aligned sequences	Grep_as
Paste_greps	Puts the alignment coordinates and statistics on one line	Paste_greps
Get_albcsi	Calls reformat_grep for all of the pasted .grep files	Get_albcsi
Coord_extr	Calls ext_coord for all files with the extension; .uniq	Coord_extr
Sort_by_len	Calls Bins, puts in upper and lower limits for a length	Sort_by_len
Sort_by_bc	Calls Bins2, puts in upper and lower limits for b+c length	Sort_by_bc
put_in_bins	Calls Bins3, puts in upper and lower limits for % identity	put_in_bins

[a]Files are viewable at http://dir.niehs.nih.gov/ALU/methods.html except for the C programs, which were written, and are maintained and are available at the Genetic Information Research Institute.
[b]These programs request parameters from inside the program.
[c]The filenames cc, dc, and dd may be substituted for cd.
[d]Extracts sequences from the 1st *Alu* in pair, 2nd *Alu* in the pair, and the entire region of the pair and renames them $1.nd, $1.st, and $1.tot respectively ($1 = [input file]).
[e]The file extensions .nd and .tot may be substituted for .st.

needed to bring two sequences into alignment. The similarity score = (total matches) / (total matches + total mismatches + total "num(ber) of gaps"); "percent identity" = (matches) / (alignment length) (note that the similarity score and identity are approximately equal because most gaps are one or a few bases).

The "spacer len(gth) (c) " = (*Alu*2 fragment start) − (*Alu*1 fragment finish). Flanking sequences were identified that corresponded to sequences at either end of an *Alu* (or fragment) that were not found in the other member of the pair (note that the *Alu*s were initially identified in terms of a consensus *Alu*). These most likely were the result of one *Alu* being truncated relative to the other. An internal flank is referred to as "b" so that the distance between aligned regions is actually b + c.

We next recovered the sequence description headings and the original unaligned sequence fragment coordinates from the sequence files that were used in generating the data. Once the files containing the coordinates for first and second *Alu*s in a pair and the description were generated, they were pasted side-by-side with the "grepped" alignment statistics.

A Perl script called `Bins 3` was used to subdivide the summary output from the alignment into different groupings dependent on the percentage identity between the two *Alu*s in the pairs. In the process, the `Bins` program verified that the loci names in the pasted coordinate data matched the align-

ment data. The `Bins` program also recalculated the "c" size by subtracting the end of the first original *Alu* fragment coordinate from the end of the second original *Alu* fragment coordinate as an internal check.

Another Perl program, `Bins`, was written to subdivide the files according to the range in which the "a" length fell to group the data based on length. The data also were grouped according to the percentage identity between the two *Alu* sequences in a pair. Finally, two more Perl programs were used to subdivide the data according to distances between the *Alu*s (`Bins C`) and distances between aligned regions (`Bins 2`).

The data were summarized in four hypertext mark-up language (HTML) tables (cc.html, dd.html, etc.) These tables include the raw data supporting the present results. Links from each of the cells in the tables contain all *Alu* sequence information available as of September 1999. Each cell has a hypertext link to the characteristics describing each pair. An example of the linked information is provided in Figure 2.

Determination of Internal Truncations

It was necessary to use four different algorithms, one for each of the four possible orientations to determine the length of internal orientations. In the case of CC, p is equal to the fragment finish coordinate minus the alignment finish coordinate of the first *Alu* in the pair, while q equals the fragment finish coordinate minus the alignment finish coordinate of the second *Alu* in the pair. If $q - p > 1$, then the first *Alu* head (internal in the pair) is truncated relative to the second. Otherwise, if they are not equal, the second head (external) is truncated.

For DD, the algorithm is as follows: p is equal to the aligned start coordinate minus the fragment start coordinate of the first *Alu* in the pair, while q equals the aligned start coordinate minus the fragment start coordinate of the second *Alu* in the pair. If $q - p > 1$, then the first *Alu* head (external in the pair) is truncated relative to the second. Otherwise, if $p - q > 1$, then the second *Alu* head (internal head in the pair) is truncated relative to the first.

Inverted repeats were more complicated. For CD, where both heads are internal, p is equal to the fragment finish coordinate minus the aligned finish coordinate of the first *Alu* in the pair, while q equals the aligned start coordinate minus the fragment start coordinate of the second *Alu* in the pair. If $q - p > 1$, then the first *Alu* head is truncated relative to the second. Otherwise, if $p - q > 1$, then the second *Alu* head is truncated relative to the first.

Finally, for DC, where both heads are external, p is equal to the aligned start coordinate minus the fragment start coordinate of the first *Alu* in the pair, while q equals the fragment finish coordinate minus the alignment finish coordinate of the second *Alu* in the pair. If $q - p > 1$, then the first *Alu* head is truncated relative to the second. Otherwise, if $p - q > 1$, then the second *Alu* head is truncated relative to the first.

ACKNOWLEDGMENTS

We are grateful to Paul Klonowski for writing `PRENUM02` and for automating the tabulation of the data in HTML. We greatly appreciate the comments of Rob Slebbos and Jim Mason on the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Akgun, E., Zahn, J., Baumes, S., Brown, G., Liang, F., Romanienko, P.J., Lewis, S., and Jasin, M. 1997. Palindrome resolution and recombination in the mammalian germ line. *Mol. Cell. Biol.* **17:** 5559–5570.

Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., and Zoghbi, H.Y. 1999. Rett syndrome is caused by mutations in X-linked MECP2, encoding metyl-CpG-binding protein 2. *Nat. Genet.* **23:** 185–188.

Anderson, J.L., Carlquist, J.F., King, G.J., Morrison, L., Thomson, M.J., Ludwig, E.H., Muhlestein, J.B., Bair, T.L., and Ward, R.H. 1998. Angiotensin-converting enzyme genotypes and risk for myocardial infarction in women. *J. Am. Coll. Cardiol.* **31:** 790–796.

Arcot, S.S., Shaikh, T.H., Kim, J., Bennett, L., Alegria-Hartman, M., Nelson, D.O., Deininger, P.L., and Batzer, M.A. 1995. Sequence diversity and chromosomal distribution of "young" *Alu* repeats. *Gene* **163:** 273–278.

Aronin, N., Chase, K., Young, C., Sapp, E., Schwarz, C., Matta, N., Kornreich, R., Landwehrmeyer, B., Bird, E., Beal, M.F., et al. 1995. CAG expansion affects the expression of mutant Huntingtin in the Huntington's disease brain. *Neuron* **15:** 1193–1201.

Batzer, M.A., Arcot, S.S., Phinney, J.W., Alegria-Hartman, M., Kass, D.H., Milligan, S.M., Kimpton, C., Gill, P., Hochmeister, M., Ioannou, P.A., et al. 1996. Genetic variation of recent *Alu* insertions in human populations. *J. Mol. Evol.* **42:** 22–29.

Burch, G.H., Gong, Y., Liu, W., Dettman, R.W., Curry, C.J., Smith, L., Miller, W.L., and Bristow, J. 1997. Tenascin-X deficiency is associated with Ehlers-Danlos syndrome. *Nat. Genet.* **17:** 104–108.

Chae, J.J., Park, Y.B., Kim, S.H., Hong, S.S., Song, G.J., Han, K.H., Namkoong, Y., Kim, H.S., and Lee, C.C. 1997. Two partial deletion mutations involving the same *Alu* sequence within intron 8 of the LDL receptor gene in Korean patients with familial hypercholesterolemia. *Hum. Genet.* **99:** 155–163.

Chen, S.J., Chen, Z., d'Auriol, L., M. Le Coniat, Grausz, D., and Berger, R. 1989. Ph1 +bcr- acute leukemias: implication of *Alu* sequences in a chromosomal translocation occuring in the new cluster region within the BCR gene. *Oncogene* **4:** 195–202.

Chesnokov, I., Chu, W.M., Botchan, M.R., and Schmid, C.W. 1996. p53 inhibits RNA polymerase III-directed transcription in a promoter-dependent manner. *Mol. Cell Biol.* **16:** 7084–7088.

Chu, W.M., Ballard, R., Carpick, B.W., Williams, B.R., and Schmid, C.W. 1998. Potential *Alu* function: Regulation of the activity of double-stranded RNA-activated kinase PKR. *Mol. Cell Biol.* **18:** 58–68.

Deininger, P.L. and Batzer, M.A. 1999. *Alu* repeats and human disease. *Mol. Genet. Metabol.* **67:** 183–193.

Derr, L.K. and Strathern, J.N. 1993. A role for reverse transcripts in gene conversion. *Nature* **361:** 170–173.

Erlich, D.S. 1989. Illegitimate recombination in bacteria. In *Illegitimate recombination in bacteria*, (eds. D.E. Berg and M.M. Howe), pp. 799–832. American Society for Microbiology, Washington D.C.

Faulkner, D.V. and Jurka, J. 1988. Multiple aligned sequence editor (MASE). *Trends Biochem. Sci.* **13:** 321–322.

Frangiskakis, J.M., Ewart, A.K., Morris, C.A., Mervis, C.B., Bertrand, J., Robinson, B.F., Klein, B.P., Ensing, G.J., Everett, L.A., Green, E.D., et al. 1996. LIM-kinase1 hemizygosity implicated in impaired visuospatial constructive cognition. *Cell* **86:** 59–69.

Gebow, D., Miselis, N., and Liber, H.L. 2000. Homologous and nonhomologous recombination resulting in deletion: effects of p53 status, microhomology, and repetitive DNA length and orientation. *Mol. Cell Biol.* **20:** 4028–4035.

Giordano, R., Magnano, A.R., Zaccagnini, G., Pittoggi, C., Moscufo, N., Lorenzini, R., and Spadafora, C. 2000. Reverse transcriptase

activity in mature spermatozoa of mouse. *J. Cell Biol.* **148:** 1107–1113.
Gordenin, D.A., Kunkel, T.A., and Resnick, M.A. 1997. Repeat expansion—all in a flap? *Nat. Genet.* **16:** 116–118.
Gordenin, D.A., Lobachev, K.S., Degtyareva, N.P., Malkova, A.L., Perkins, E., and Resnick, M.A. 1993. Inverted DNA repeats: A source of eukaryotic genomic instability. *Mol. Cell Biol.* **13:** 5315–5322.
Gordenin, D.A. and Resnick, M.A. 1998. Yeast ARMs (DNA at-risk motifs) can reveal sources of genome instability. *Mutat. Res.* **400:** 45–58.
Gu, J.J., Spychala, J., and Mitchell, B.S. 1997. Regulation of the human inosine monophosphate dehydrogenase type I gene. Utilization of alternative promoters. *J. Biol. Chem.* **272:** 4458–4466
Harteveld, K.L., Losekoot, M., Fodde, R., Giordano, P.C., and Bernini, L.F. 1997. The involvement of *Alu* repeats in recombination events at the α-globin gene cluster: Characterization of two *alphazero-thalassaemia* deletion breakpoints. *Hum. Genet.* **99:** 528–534.
Heikkinen, J., Toppinen, T., Yeowell, H., Krieg, T., Steinmann, B., Kivirikko, K.I. and Myllyla, R. 1997. Duplication of seven exons in the lysyl hydroxylase gene is associated with longer forms of a repetitive sequence within the gene and is a common cause for the type VI variant of Ehlers-Danlos syndrome. *Am. J. Hum. Genet.* **60:** 48–56.
Jurka, J. 1993. A new subfamily of recently retroposed human *Alu* repeats. *Nucleic Acids Res.* **21:** 2252.
———. 1995. Origin and evolution of *Alu* repetitive elements. In *Origin and evolution of* Alu *repetitive elements* (ed. Maraia, R.J.), pp. 25–42. Springer, New York.
———. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci.* **94:** 1872–1877.
———. 1998. Repeats in genomic DNA: Mining and meaning. *Curr. Opin. Struct. Biol.* **8:** 333–337.
Kitamura, H., Moriyama, T., Izumi, M., Yokoyama, K., Yamauchi, A., Ueda, N., Kamada, T., and Imai, E. 1996. Angiotensin I-converting enzyme insertion/deletion polymorphism: Potential significance in nephrology. *Kidney Int. Suppl.* **55:** S101–S103.
Klonowski, P. 1998. PRENUM02. Genetic Information Research Institute, Sunnyvale, CA
Klonowski, P. and Jurka, J. 1997. VEXT. Genetic Information Research Institute, Sunnyvale, CA
Leach, D.R. 1994. Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *Bioessays* **16:** 893–900.
Lewis, S., Akgun, E., and Jasin, M. 1999. Palindromic DNA and genome stability. Further studies. *Ann. NY Acad. Sci.* **870:** 45–57.
Lewis, S.M. 1999. Palindromy is eliminated through a structure-specific recombination process in rodent cells. *Nucleic Acids Res.* **27:** 2521–2528.
Lobachev, K.S., Shor, B.M., Tran, H.T., Taylor, W., Keen, J.D., Resnick, M.A., and Gordenin, D.A. 1998. Factors affecting inverted repeat stimulation of recombination and deletion in *Saccharomyces cerevisiae*. *Genetics* **148:** 1507–1524.
Lobachev, K.S., Stenger, J.E., Kozyreva, O.G., Jurka, J., Gordenin, D.A., and Resnick, M.A. 2000. Inverted *Alu* repeats unstable in yeast are excluded from the human genome. *Embo. J.* **19:** 3822–3830.
Miki, Y., Katagiri, T., Kasumi, F., Yoshimoto, T., and Nakamura, Y. 1996. Mutation analysis in the BRCA2 gene in primary breast cancers. *Nat. Genet.* **13:** 245–247.
Mullersman, J.E. and Pfeffer, L.M. 1995. An *Alu* cassette in the cytoplasmic domain of an interferon receptor subunit. *J. Interferon Cytokine Res.* **15:** 815–817.
Muratani, K., Hada, T., and Higashino, K. 1993. Gene analysis of human cholinesterase variants. *Nippon Rinsho* **51:** 495–500.
Myerowitz, R. and Hogikyan, N.D. 1987. A deletion involving *Alu* sequences in the β-hexosaminidase α-chain gene of French Canadians with Tay-Sachs disease. *J. Biol. Chem.* **262:** 15396–15399.
Nag, D.K. and Kurst, A. 1997. A 140-bp-long palindromic sequence induces double-strand breaks during meiosis in the yeast *Saccharomyces cerevisiae*. *Genetics* **146:** 835–847.
Norris, J., Fan, D., Aleman, C., Marks, J.R., Futreal, P.A., Wiseman, R.W., Iglehart, J.D., Deininger, P.L., and McDonnell, D.P. 1995. Identification of a new subclass of *Alu* DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *J. Biol. Chem.* **270:** 22777–22782.
Novick, G.E., Batzer, M.A., Deininger, P.L., and Herrers, R.J. 1996. The mobile genetic element *Alu* in the human genome. *BioScience* **46:** 32–41.
Pousi, B., Hautala, T., Heikkinen, J., Pajunen, L., Kivirikko, K.I., and Myllyla, R. 1994. *Alu-Alu* recombination results in a duplication of seven exons in the lysyl hydroxylase gene in a patient with the type VI variant of Ehlers-Danlos syndrome. *Am. J. Hum. Genet.* **55:** 899–906.
Pousi, B., Hautala, T., Hyland, J.C., Schroter, J., Eckes, B., Kivirikko, K.I., and Myllyla, R. 1998. A compound heterozygote patient with Ehlers-Danlos syndrome type VI has a deletion in one allele and a splicing defect in the other allele of the lysyl hydroxylase gene. *Hum. Mutat.* **11:** 55–61.
Resnick, M.A. and Cox, B.S. 2000. Yeast as an honorary mammal. *Mutat. Res.* **451:** 1–11.
Robinson, W.P., Waslynka, J., Bernasconi, F., Wang, M., Clark, S., Kotzot, D., and Schinzel, A. 1996. Delineation of 7q11.2 deletions associated with Williams-Beuren syndrome and mapping of a repetitive sequence to within and to either side of the common deletion. *Genomics* **34:** 17–23.
Ruskin, B. and Fink, G.R. 1993. Mutations in POL1 increase the mitotic instability of tandem inverted repeats in *Saccharomyces cerevisiae*. *Genetics* **134:** 43–56.
Shen, M., Batzer, M., and Deininger, P. 1991. Evolution of the master *Alu* gene(s). *J. Mol. Evol.* **33:** 311–320.
Slebos, R.J., Resnick, M.A., and Taylor, J.A. 1998. Inactivation of the p53 tumor suppressor gene via a novel *Alu* rearrrangement. *Cancer Res.* **58:** 5333–5336.
Strout, M.P., Marcucci, G., Bloomfield, C.D., and Caligiuri, M.A. 1998. The partial tandem duplication of ALL1 (MLL) is consistently generated by *Alu*-mediated homologous recombination in acute myeloid leukemia. *Proc. Natl. Acad. Sci.* **95:** 2390–2395.
Szmulewicz, M.N., Novick, G.E., and Herrera, R.J. 1998. Effects of *Alu* insertions on gene function. *Electrophoresis* **19:** 1260–1264.
Toriello, H.V., Glover, T.W., Takahara, K., Byers, P.H., Miller, D.E., Higgins, J.V., and Greenspan, D.S. 1996. A translocation interrupts the COL5A1 gene in a patient with Ehlers-Danlos syndrome and hypomelanosis of Ito. *Nat. Genet.* **13:** 361–365.
Tran, H., Degtyareva, N., Gordenin, D., and Resnick, M.A. 1997. Altered replication and inverted repeats induce mismatch repair-independent recombination between highly diverged DNAs in yeast. *Mol. Cel. Biol.* **17:** 1027–1036.
Tsuchiya, T., Saegusa, Y., Taira, T., Mimori, T., Iguchi-Ariga, S.M., and Ariga, H. 1998. Ku antigen binds to *Alu* family DNA. *J. Biochem. (Tokyo)* **123:** 120–127.
Vidaud, D., Vidaud, M., Bahnak, B.R., Siguret, V., Gispert Sanchez, S., Laurian, Y., Meyer, D., Goossens, M., and Lavergne, J.M. 1993. Haemophilia B due to a de novo insertion of a human-specific *Alu* subfamily member within the coding region of the factor IX gene. *Eur. J. Hum. Genet.* **1:** 30–36.
Wallace, M.R., Andersen, L.B., Saulino, A.M., Gregory, P.E., Glover, T.W,. and Collins, F.S. 1991. A de nove Alu insertion results in neurofibromatosis type1. *Nature* 864–866.
Waterman, M.S. 1984. Efficient sequence alignment algorithms. *J. Theor. Biol.* **108:** 333–337.

Received August 10, 2000; accepted in revised form November 6, 2000.